Combining Estimates from Related Surveys via Bivariate Models

(Application: using ACS estimates to improve estimates from smaller U.S. surveys)

William R. Bell and Carolina Franco, U.S. Census Bureau

2016 Ross-Royall Symposium

February 26, 2016

Disclaimer:

This report is released to inform interested parties of ongoing research and to encourage discussion. The views expressed on statistical, methodological, technical, or operational issues are those of the author(s) and not necessarily those of the U.S. Census Bureau.

Introduction

- Investigate the potential of using bivariate models to borrow strength from estimates from a large survey to improve related estimates from smaller surveys.
- Motivation: "Large survey" is the Census Bureau's American Community Survey (ACS), the largest U.S. household survey.
- Approach is simple and requires no covariates from auxiliary information.
- Real examples show that large reductions in standard errors of estimates are possible.

ACS: The Largest U.S. Household Survey

American Community Survey (ACS)

- Conducted annually (data collected throughout the year) and has replaced the decennial census long form sample.
- Samples approximately 3.5 million addresses each year.
- Encompasses a broad range of topics: demographic, income, health insurance, employment, disabilities, occupations, housing, education, veteran status, etc.
- Produces estimates annually based on 1 or 5 years of data.

Three Smaller U.S. Surveys

- **Survey of Income and Program Participation (SIPP) Disability Module**
  - Approx. 37,000 households and 70,000 persons in 2008 panel.
  - Detailed questions about many different aspects of disability.
- **National Health Interview Survey (NHIS)**
  - About 110,000 persons in Family Core component, 2013.
  - Questions about a broad range of health topics asked in personal household interviews.
  - Estimates used to track health status, health care access, and progress toward achieving national health objectives
- **Current Population Survey (CPS) Annual Social and Economic Supplement**.
  - Samples about 100,000 addresses.
  - Provides official national estimates of income and poverty.

Four Applications

1. **SIPP estimates of U.S. state disability rates**.
   ACS variable: Estimate of state disability rates (types of disabilities and the time frames differ from SIPP).
2. **NHIS estimates of U.S. state uninsured rates**.
   ACS variable: Estimate of U.S. state uninsured rates (questions asked and the mode of survey delivery and design differ from NHIS).
3. **CPS estimates of per capita expenditure on health insurance premiums by state**
   ACS variable: Estimated per capita income by state.
4. **ACS 1-yr estimates (of anything! Take county rates of children in poverty to illustrate)**
   2nd variable: Corresponding previous ACS 5-yr estimates (larger sample size, but less current).

Univariate Gaussian Shrinkage Model for Survey Estimates

- For m small areas:

$$y_i = Y_i + e_i \qquad i = 1, \ldots, m$$
$$Y_i = \mu + u_i$$

- y_i is the direct survey estimate of Y_i, the population characteristic of interest for area i.
- e_i is the sampling error in y_i, generally assumed to be $N(0, v_i)$, independent with v_i known.
- u_i is the area i random effect, usually assumed to be *i.i.d.* $N(0, \sigma_u^2)$ and independent of the e_i.

Shrinkage Estimation (Stein 1956, Carter and Rolph 1974)

- Best linear predictor of Y_i (μ and σ^2 known):

$$\hat{Y}_i = (1 - \gamma_i) y_i + \gamma_i \mu$$

where

$$\gamma_i = \frac{v_i}{v_i + \sigma_u^2}$$

- Weighted average $\hat{Y}_i$ "shrinks" the direct estimate y_i towards the overall mean μ.
- The smaller is the sampling variance v_i the more weight is placed on the direct survey estimate y_i.
- Parameters unknown: estimate by ML or REML, or take Bayesian approach.
- Fay and Herriot (1979) extended the approach to shrink y_i towards a regression mean $\mu_i = x_i' \beta$, and applied this approach to small area estimation.

Bivariate Gaussian Model

$$y_{1i} = Y_{1i} + e_{1i} = (\mu_1 + u_{1i}) + e_{1i}, \quad i = 1, \ldots, m.$$
$$y_{2i} = Y_{2i} + e_{2i} = (\mu_2 + u_{2i}) + e_{2i}$$

$$\begin{bmatrix} u_{1i} \\ u_{2i} \end{bmatrix} \overset{i.i.d}{\sim} N(0, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

$$\begin{bmatrix} e_{1i} \\ e_{2i} \end{bmatrix} \overset{i.i.d}{\sim} N(0, \mathbf{V}_i), \quad \mathbf{V}_i = \begin{bmatrix} v_{11} & 0 \\ 0 & \sigma_{22} \end{bmatrix}$$

- y_{1i} is the direct estimate of the quantity of interest Y_{1i}, and y_{2i} is the direct estimate from another survey of a related quantity Y_{2i}.
- Note that $\mathbf{V}_i$ assumes the sampling errors e_{1i} and e_{2i} are uncorrelated. This can be generalized.
- The alternative of simply including y_{2i} as a regression covariate in the model would ignore their sampling errors!

Estimation/Inference for Model Parameters

- Unknown parameters: $\mu_1, \mu_2, \sigma_{11}, \sigma_{22}$, and σ_{12} or $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$.
- Sampling variances v_{1i} and v_{2i} are treated as known (really estimated using survey microdata).
- Can estimate unknown parameters by ML or REML.
- We shall use a Bayesian approach with flat priors on $\mu_1, \mu_2, \sigma_{11} > 0, \sigma_{22} > 0$ and $\rho \in (-1, 1)$.
- Approach was implemented in JAGS.

Prediction When Model Parameters are Known

In matrix notation $\mathbf{y}_i = \mathbf{Y}_i + \mathbf{e}_i = (\boldsymbol{\mu} + \mathbf{u}_i) + \mathbf{e}_i$

- $\hat{\mathbf{Y}}_i^{BP} = E(\mathbf{Y}_i|\mathbf{y}_i) = \boldsymbol{\mu} + \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \mathbf{V}_i)^{-1}(\mathbf{y}_i - \boldsymbol{\mu})$
- $MSE(\hat{\mathbf{Y}}_i^{BP}) = Var(\mathbf{Y}_i|\mathbf{y}_i) = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \mathbf{V}_i)^{-1}\boldsymbol{\Sigma}$
- We are interested in predicting Y_{1i} only, not Y_{2i}
- $\hat{Y}_{1i}^{BP}$ is a linear combination of μ_1, $(y_{1i} - \mu_1)$, and $(y_{2i} - \mu_2)$.

MSE % Reductions from Shrinkage Estimation

- direct estimation to univariate shrinkage:

$$100 \times \left\{ 1 - \frac{\text{Var}(Y_{1i}|y_{1i})}{v_{1i}} \right\}$$

(more reduction as v_{1i} increases)

MSE % Reductions from Shrinkage Estimation

- direct estimation to univariate shrinkage:

$$100 \times \left\{ 1 - \frac{\text{Var}(Y_{1i}|y_{1i})}{v_{1i}} \right\}$$

(more reduction as v_{1i} increases)

- univariate to bivariate shrinkage:

$$100 \times \left\{ 1 - \frac{\text{Var}(Y_{1i}|y_{1i}, y_{2i})}{\text{Var}(Y_{1i}|y_{1i})} \right\}$$

(more reduction as v_{2i} decreases and as ρ increases)

MSE % Reductions from Shrinkage Estimation

- direct estimation to univariate shrinkage:

$$100 \times \left\{ 1 - \frac{\text{Var}(Y_{1i}|y_{1i})}{v_{1i}} \right\}$$

  (more reduction as v_{1i} increases)

- univariate to bivariate shrinkage:

$$100 \times \left\{ 1 - \frac{\text{Var}(Y_{1i}|y_{1i}, y_{2i})}{\text{Var}(Y_{1i}|y_{1i})} \right\}$$

  (more reduction as v_{2i} decreases and as ρ increases)

- direct estimation to bivariate shrinkage:

$$100 \times \left\{ 1 - \frac{\text{Var}(Y_{1i}|y_{1i}, y_{2i})}{v_{1i}} \right\}$$

Application I: 2010 Disability Rates for U.S. States: SIPP borrowing from ACS

y_{1i} = SIPP disability estimate, y_{2i} = ACS disability estimate

Smoothing of SIPP direct sampling variance estimates is applied.

$\hat{\rho} = .\mathbf{82}$

- Univariate shrinkage yields an MSE decrease of $2\% - 67\%$ from direct, with a median of 19%

Application I: 2010 Disability Rates for U.S. States: SIPP borrowing from ACS

y_{1i} = SIPP disability estimate, y_{2i} = ACS disability estimate

Smoothing of SIPP direct sampling variance estimates is applied.

$\hat{\rho} = .\mathbf{82}$

- Univariate shrinkage yields an MSE decrease of $2\% - 67\%$ from direct, with a median of 19%
- The MSE decrease from bivariate vs. univariate model is $6\% - 59\%$ with a median of 29%

Application I: 2010 Disability Rates for U.S. States: SIPP borrowing from ACS

y_{1i} = SIPP disability estimate, y_{2i} = ACS disability estimate

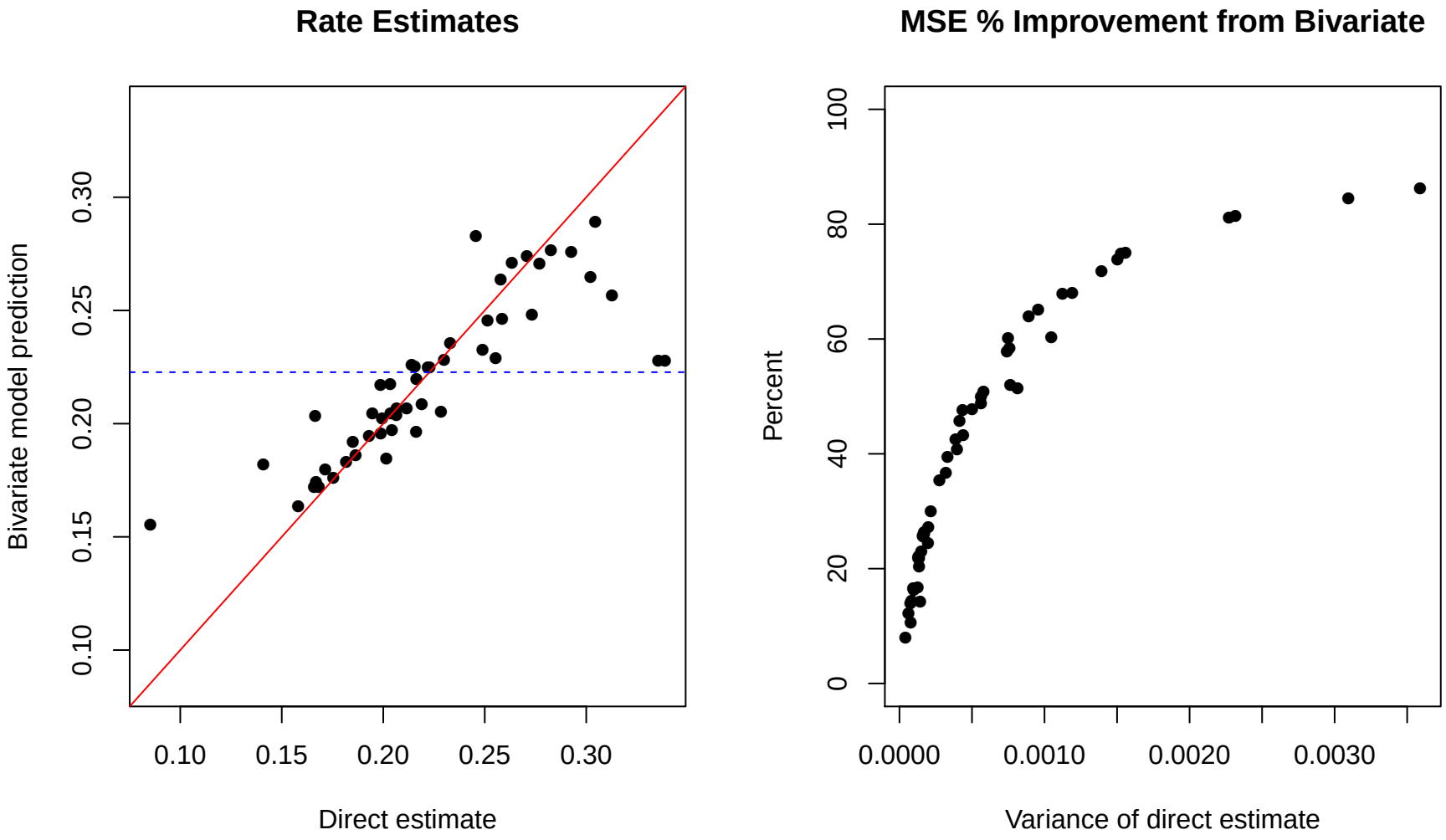
Smoothing of SIPP direct sampling variance estimates is applied.

$\hat{\rho}$= **.82**

- Univariate shrinkage yields an MSE decrease of $2\% - 67\%$ from direct, with a median of 19%
- The MSE decrease from bivariate vs. univariate model is $6\% - 59\%$ with a median of 29%
- The MSE decrease from bivariate vs. direct is $\mathbf{8 - 86\%}$, **with a median decrease of 43%**

Disability Rates for U.S. States, 2014

Bivariate model for SIPP and ACS estimates



Application II: 2013 Health Insurance Coverage Rates for U.S. States: NHIS Borrowing from ACS

y_{1i} = NHIS estimate of health insurance coverage (from National Center for Health Statistics)
y_{2i} = ACS estimate of health insurance coverage

Estimates published for only 43 states "due to considerations of sample size and precision."

$\hat{\rho} = \mathbf{.96}$

- MSE decrease UNI vs. Direct: $1\% - 16\%$, median $= 10\%$

Using bivariate model might allow publication of estimates for states that would otherwise be excluded (?)

Application II: 2013 Health Insurance Coverage Rates for U.S. States: NHIS Borrowing from ACS

y_{1i} = NHIS estimate of health insurance coverage (from National Center for Health Statistics)
y_{2i} = ACS estimate of health insurance coverage

Estimates published for only 43 states "due to considerations of sample size and precision."

$\hat{\rho}$= **.96**

- MSE decrease UNI vs. Direct: $1\% - 16\%$, median $= 10\%$
- MSE decrease BIV vs. UNI: $16\% - 67\%$, median $= 54\%$

Using bivariate model might allow publication of estimates for states that would otherwise be excluded (?)

Application II: 2013 Health Insurance Coverage Rates for U.S. States: NHIS Borrowing from ACS

y_{1i} = NHIS estimate of health insurance coverage (from National Center for Health Statistics)
y_{2i} = ACS estimate of health insurance coverage

Estimates published for only 43 states "due to considerations of sample size and precision."

$\hat{\rho}$= **.96**

- MSE decrease UNI vs. Direct: $1\% - 16\%$, median $= 10\%$
- MSE decrease BIV vs. UNI: $16\% - 67\%$, median $= 54\%$
- MSE **decrease BIV vs. Direct: 19 – 72%, median = 60%**!

Using bivariate model might allow publication of estimates for states that would otherwise be excluded (?)

Application III: 2012 Per Capita Expenditures for Health Insurance for U.S. States: CPS Borrowing from ACS

y_{1i} = CPS estimated per capita expenditure on health insurance premimums
y_{2i} = ACS per capita income estimate

$\hat{\rho}$= **.65**

- MSE decrease UNI vs. Direct: $1\% - 55\%$, median $= 8\%$

More modest decreases overall, presumably because ρ and v_{1i}/σ_{11} are lower than in the previous examples.

Application III: 2012 Per Capita Expenditures for Health Insurance for U.S. States: CPS Borrowing from ACS

y_{1i} = CPS estimated per capita expenditure on health insurance premimums
y_{2i} = ACS per capita income estimate

$\hat{\rho} = \mathbf{.65}$

- MSE decrease UNI vs. Direct: $1\% - 55\%$, median $= 8\%$
- MSE decrease BIV vs. UNI: $-1.5\% - 28\%$, median $= 6\%$

More modest decreases overall, presumably because ρ and v_{1i}/σ_{11} are lower than in the previous examples.

Application III: 2012 Per Capita Expenditures for Health Insurance for U.S. States: CPS Borrowing from ACS

y_{1i} = CPS estimated per capita expenditure on health insurance premimums
y_{2i} = ACS per capita income estimate

$\hat{\rho}$= **.65**

- MSE decrease UNI vs. Direct: $1\% - 55\%$, median $= 8\%$
- MSE decrease BIV vs. UNI: $-1.5\% - 28\%$, median $= 6\%$
- MSE decrease BIV vs. Direct: $2\% - 68\%$, median $= 14\%$

More modest decreases overall, presumably because ρ and v_{1i}/σ_{11} are lower than in the previous examples.

Application IV: ACS 1-yr County Poverty Estimates Borrow from Previous ACS 5-yr County Poverty Estimates

y_{1i} = 2012 ACS estimated county rates of children in poverty

y_{2i} = 2007-2011 ACS estimated county child poverty rates

Note: Good covariates are available for modeling, but are not used here.

$\hat{\rho} = .\mathbf{94}$

- MSE decrease UNI vs. Direct: $0.4\% - 87\%$, median $= 32\%$

Application IV: ACS 1-yr County Poverty Estimates Borrow from Previous ACS 5-yr County Poverty Estimates

y_{1i} = 2012 ACS estimated county rates of children in poverty

y_{2i} = 2007-2011 ACS estimated county child poverty rates

Note: Good covariates are available for modeling, but are not used here.

$\hat{\rho} = .\mathbf{94}$

- MSE decrease UNI vs. Direct: $0.4\% - 87\%$, median $= 32\%$
- MSE decrease BIV vs. UNI: $4\% - 65\%$, median $= 49\%$

Application IV: ACS 1-yr County Poverty Estimates Borrow from Previous ACS 5-yr County Poverty Estimates

y_{1i} = 2012 ACS estimated county rates of children in poverty

y_{2i} = 2007-2011 ACS estimated county child poverty rates

Note: Good covariates are available for modeling, but are not used here.

$\hat{\rho} = .\mathbf{94}$

- MSE decrease UNI vs. Direct: $0.4\% - 87\%$, median $= 32\%$
- MSE decrease BIV vs. UNI: $4\% - 65\%$, median $= 49\%$
- **MSE decrease BIV vs. Direct: $4 - 91\%$, median $= 67\%$!!**

Concluding Remarks

- Bivariate model can achieve large MSE decreases by borrowing strength from ACS estimates to improve estimates from smaller surveys, provided ρ is high!
- Model is simple; key is the quality of the additional data source (ACS estimates) used for this purpose.
- In most of the examples (I, II, IV), the biggest part of the MSE decreases came from the univariate to bivariate shrinkage, not from the univariate shrinkage.
- Theoretical and empirical results show not much improvement when a larger survey borrows strength from a smaller one.